



Managed by Fermi Research Alliance, LLC for the U.S. Department of Energy Office of Science

DATA INTENSIVE SCIENTIFIC WORKFLOWS ON A FEDERATED CLOUD

Cooperative Research and Development Final Report

CRADA Number: FRA-2015-0001

Fermilab Technical Contact: Gabriele Garzoglio

Summary Report
31 October 2015

NOTICE

This report was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or any agency thereof.

Available electronically at <http://www.osti.gov/bridge>

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:
U.S. Department of Energy Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
phone: 865.576.8401
fax: 865.576.5728
email: <mailto:reports@adonis.osti.gov>

Available for sale to the public, in paper, from:
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
phone: 800.553.6847
fax: 703.605.6900
email: orders@ntis.fedworld.gov
online ordering: <http://www.ntis.gov/ordering.htm>

In accordance with Requirements set forth in Article XI.A(3) of the CRADA document, this document is the final CRADA report, including a list of Subject Inventions, to be forwarded to the Office of Science and Technical Information as part of the commitment to the public to demonstrate results of federally funded research.

CRADA Number: FRA 2015-0001

CRADA Title: DATA INTENSIVE SCIENTIFIC WORKFLOWS ON A FEDERATED CLOUD

Parties to the Agreement: Korean Institute of Science and Technology Information (KISTI) and Fermi Research Alliance

Abstract of CRADA work:

The Fermilab Scientific Computing Division and the KISTI Global Science Experimental Data Hub Center have built a prototypical large-scale infrastructure to handle scientific workflows of stakeholders to run on multiple cloud resources. The demonstrations have been in the areas of (a) Data-Intensive Scientific Workflows on Federated Clouds, (b) Interoperability and Federation of Cloud Resources, and (c) Virtual Infrastructure Automation to enable On-Demand Services.

Summary of Research Results:

The work resulted in demonstrations and studies to build a production scale infrastructure to run scientific workflows on dynamically provisioned resources. The results are organized in the three major areas below.

1. Data-Intensive Scientific Workflows on Federated Clouds focused on developing and integrating mechanisms to support the execution of scientific workflows with large data processing needs. The deliverables for this area are the following:

- a. Cost-sensitive provisioning on the AWS spot market: focus previous studies on cost-sensitive provisioning algorithms to identify optimal bidding strategies to provision resources on the AWS spot market. Hao Wu, a PhD student from the Illinois Institute of Technology has focused his research on this topic. The results of his research this year have been accepted at the MTAGS workshop as a paper jointly authored with KISTI.
- b. Investigate execution of workflows for CERN LHC experiments: the HEP Cloud Facility has run Monte Carlo workflows on AWS and Fermilab resources at the scale of 56,000 cores (25% of global capacity) for one month in February 2016. A demonstration at Supercomputing 2015 has shown the capability of gCloud at KISTI to be integrated with that activity.
- c. Integration of RnD infrastructure with the Fermilab HEPCloud Facility: run scientific workflows on federated cloud resources via the GlideinWMS system and cloud web

17 December 2014

services API's. This activity demonstrated that workflows for the CMS and NOvA experiments can take advantage of Cloud resources for their computational peaks through the HEPCloud Facility infrastructure.

2. **Interoperability and Federation of Cloud Resources** found a set of virtual image formats and application programming interfaces that can be used by all members of a virtual organization across a heterogeneous infrastructure. The deliverables for this area are the following:
 - a. Improve data management by developing strategies to interact with Amazon Simple Storage Service (S3) effectively: develop tools to use S3 as a Storage Element fully integrated with the experiments data management services.
 - b. Demonstrate a federated Cloud between Fermilab and KISTI: a demonstration at SC2015 has shown the federation capabilities between the HEPCloud Facility at FNAL and gCloud at KISTI, running CMS Monte Carlo workflows.
 - c. VM image portability: improve AWS authentication mechanism of the automatic virtual machine image format conversion tool developed in 2014.
 - d. Perform benchmarks of Fermilab and AWS computing infrastructure: the results are used to scale the expected execution times and evaluate costs more accurately.
3. **Virtual Infrastructure Automation for On-demand Services** found the most efficient methods for scientific grid and cloud computing middleware to distribute data and execution across the WAN to meet the demand. The deliverables for this area are the following:
 - a. Provisioning of a platform of services: transition the mechanism to provision complicated ensembles of virtual machines in support of scientific workflows to use native AWS orchestration services (CloudFormation). This mechanism was applied to web caching services and provides automatic service discovery and scaling on demand.
 - b. Improve the accounting and monitoring infrastructure: AWS provides accounting infrastructure to track cost by VM and services used; the additional infrastructure developed by this CRADA allows to compare user job execution time with overall VM running time and associate costs directly with scientific workflow execution

Subject Inventions listing: None

Report Date: 10/31/2015

Technical Contact at Fermilab: Gabriele Garzoglio

This document contains NO confidential, protectable or proprietary information.

31 October 2015